IN THE CLAIMS

Please amend claims as indicated below:

1.   (Previously Presented) A method of categorizing an initial collection of documents, each document being represented by a string of characters, the method comprising the steps of:

      identifying predefined characters in the string of characters from the documents in the initial collection of documents to form identified characters;

      changing the identified characters in the documents in the initial collection of documents to form a preprocessed collection of documents, each of the preprocessed collection of documents represented by a preprocessed string of characters;

      constructing a number of categories from the preprocessed string of characters of the preprocessed collection of documents; and

      assigning each document in the preprocessed collection of documents to a category to form a hierarchy of categories of documents.

2.   (Previously Presented) The method of claim 1 wherein the step of constructing a number of categories includes the steps of:

      clearing a temporary category and selecting a seed document from the preprocessed collection of documents as a first document of the temporary category;

      collecting documents from the preprocessed collection of documents that are similar to the seed document into the temporary category;

      testing to determine if there are enough documents in the temporary category to merit construction of a new category;

      constructing the new category from the temporary category and generating a heading for the new category if there are enough documents in the temporary category to merit construction and generation;

      assigning the seed document to a category reserved for documents not belonging to any specific category if there are not enough documents in the temporary category; and

      marking the documents assigned to any category in the preprocessed collection of documents as processed.

3.   (Original) The method of claim 2 wherein the predefined characters include punctuation marks, and the changing step removes the punctuation marks from the string of characters.

4.   (Original) The method of claim 2 wherein the predefined characters include upper-case characters, and the changing step replaces upper-case characters with lower-case characters.

5.   (Original) The method of claim 2 wherein the predefined characters include non-root words, and the changing step replaces the non-root words with root words.

6.   (Original) The method of claim 2 wherein the predefined characters include abbreviations, and the changing step replaces the abbreviations with original words.

7.   (Original) The method of claim 2 wherein the predefined characters include articles, and the changing step removes the articles from the string of characters.

8.   (Previously Presented) The method of claim 2 wherein the collecting step further includes the step of loading a character string from the preprocessed string of characters of the seed document into a memory location to initialize values of a number of category properties for the temporary category.

9.   (Previously Presented) The method of claim 8 and further comprising the steps of:
    determining if there are documents in the preprocessed collection of documents that have not been processed with respect to the temporary category;
    if there are documents in the preprocessed collection of documents that have not been processed with respect to the temporary category, selecting a next document from the preprocessed collection of documents and measuring a similarity of the preprocessed string of characters of the next document using a similarity test between the next document and the values of the number of current category properties;
    including the next document in the temporary category if the next document passes the similarity test;
    updating the values of the number of category properties of the temporary category when the next document is included; and
    rejecting the next document if the next document fails the similarity test.

09/844,040 to Kim, et al.   3   Atty Dkt: 053684-0300105

PAGE 4/16 * RCVD AT 6/22/2004 4:00:05 PM [Eastern Daylight Time] * SVR:USPTO-EFXRF-1/1 * DNIS:8729306 * CSID:+6502234747 * DURATION (mm-ss):06-28

10.   (Previously Presented) The method of claim 9 and further comprising the step of repeating the steps of claim 9 for all documents in the preprocessed collection of documents.

11.   (Original) The method of claim 2 wherein the collecting step further includes the step of collecting more similar documents from a number of existing categories.

12.   (Previously Presented) The method of claim 11 and further comprising the steps of:
   determining if there are more documents in a number of existing categories that have not been processed with respect to the temporary category;
   if there are documents in the number of existing categories that have not been processed with respect to the temporary category, selecting a next document from the number of existing categories as a selected document and measuring a similarity of the preprocessed string of characters of the selected document using a similarity test between the selected document and values of a number of current category properties;
   including the selected document in the temporary category if the selected document passes the similarity test; and
   rejecting the selected document if the selected document fails the similarity test.

13.   (Original) The method of claim 12 and further comprising the step of repeating the steps of claim 12 for all documents in the number of existing categories.

14.   (Previously Presented) The method of claim 8 wherein the number of category properties includes a string of characters selected from the group consisting of a longest common sub-string in the title, a longest common substring in the body; and a document type index measured as list of fractional numbers for each document type.

15.   (Original) The method of claim 14 wherein a document type includes types selected from the group consisting of news article, technical documents, and poems.

16.   (Original) The method of claim 2 and further comprising the steps of:
   making sub-categories if there are too many documents in a given category; and
   post-processing the number of categorized lists of documents.

17.   (Cancelled).

18.   (Original) The method of claim 2 wherein the seed document is a first document in the preprocessed collection of documents.

19.   (Original) The method of claim 2 wherein the seed document is a document with a highest rank value among the documents not marked as processed in the preprocessed collection of documents.

20.   (Original) The method of claim 2 wherein the temporary category is tested to determine if there are enough documents in the temporary category to merit construction of a new category by accumulating the weight of each document when each document can contribute uniform weight or different weight based on the rank value of each document with higher ranked document given more weight.

21.   (Original) The method of claim 2 wherein the heading is a longest common substring in a title.

22.   (Original) The method of claim 21 wherein the heading includes a number of longest common substrings.

23.   (Original) The method of claim 1 and further comprising the steps of:
       determining if an anchor-text character string is available for the documents in the initial
              collection of documents; and
       attaching an anchor-text character string to the string of characters that represents the
              documents in the initial collection of documents when the anchor-text character
              string is available.

24.   (Original) The method of claim 23 wherein the anchor-text character string is a text used most frequently by hypertext documents.

25.   (Original) The method of claim 23 wherein the anchor-text character string is a text with a highest partial extrinsic rank value.

26. (Original) A method of categorizing an initial collection of documents, each document being represented by a string of characters, the method comprising the steps of:

constructing a number of categories from the initial collection of documents wherein a category is constructed by:

clearing a temporary category and selecting a seed document as a first document of a temporary category;

collecting documents from the initial collection of documents to the temporary category that are similar to the seed document;

testing to determine if there are enough documents in the temporary category to merit construction of a new category;

constructing the new category and generating a heading for the new category if there are enough documents in the temporary category to merit construction;

assigning the seed document to a category reserved for documents not belonging to any specific category if there are not enough documents in the temporary category; and

marking the documents assigned to any category in the initial collection of documents as processed; and

assigning each document in the initial collection of documents to a category to form a hierarchy of categories of documents.

27. (Original) The method of claim 26 wherein the collecting step further includes the step of loading a character string from the seed document into a memory location to initialize values of a number of category properties for the temporary category.

28. (Original) The method of claim 27 and further comprising the steps of:

determining if there are documents in the initial collection of documents that have not been marked as processed;

if there are documents in the initial collection of documents that have not been marked as processed, selecting a next document from the initial collection of documents and measuring a similarity with a similarity test between the selected document and a number of current category properties;

including the selected document in the temporary category if the selected document
    passes the similarity test; and

rejecting the selected document if the selected document fails the similarity test.

29.   (Original) The method of claim 28 and further comprising the step of repeating the steps
of claim 28 for all documents in initial collection of documents.

30.   (Original) The method of claim 26 wherein the collecting step further includes the step of
collecting more similar documents from a number of existing categories.

31.   (Original) The method of claim 30 and further comprising the steps of:

determining if there are more documents in the number of existing categories that have
    not been processed with respect to the temporary category;

if there are documents in the number of existing categories that have not been processed
    with respect to the temporary category, selecting a next document from the number
    of existing categories and measuring a similarity with a similarity test between the
    selected document and a number of current category properties;

including the selected document in the temporary category if the selected document
    passes the similarity test; and

rejecting the selected document if the selected document fails the similarity test.

32.   (Original) The method of claim 31 and further comprising the step of repeating the steps
of claim 31 for all documents in number of existing categories.

33.   (Original) The method of claim 1 wherein each document in the preprocessed collection
of documents is assigned to one or more categories to form a hierarchy of categories.

34.   (Original) The method of claim 26 wherein each document in the initial collection of
documents is assigned to one or more categories to form a hierarchy of categories.

35.   (Original) The method of claim 2 and further comprising the step of repeating the steps of
claim 2 until all documents in the preprocessed collection of documents are marked as assigned
to a category.

36.   (Original) The method of claim 35 wherein the documents in the preprocessed collection of documents are initialized as unmarked before selecting a first seed document.

37.   (Original) The method of claim 26 and further comprising the step of repeating the constructing steps of claim 26 until all documents in the initial collection of documents are marked as assigned to a category.

38.   (Original) The method of claim 37 wherein the documents in the preprocessed collection of documents are initialized as unmarked before selecting a first seed document.

39.   (Original) An apparatus that categorizes a collection of documents, each document being represented by a string of characters, the apparatus comprising:

   means for identifying predefined characters in the string of characters from each
      document to form identified characters;

   means for changing the identified characters in each document to form a preprocessed
      collection of documents;

   means for constructing a number of categories from the preprocessed collection of
      documents; and

   means for assigning each document in the preprocessed collection of documents to a
      category to form a number of categorized lists of documents.